2024

Pipeline Methods

BAA LTD

Dante Prisco

Dempsey Rose-Day

Yotam Shmeltzer

Phoebe Zhang

Table of Contents

Data Flow

The pipeline contains five scripts: '00_load', '01_clean', '02_preprocess', '03_compute', and '04_visualization'. These are named in run order 00 through 04. When script 04 is ran, it will automatically run all the preceding scripts. The first script, '00_load', loads in the provided datasets and sets the data types and column names. The second script, '01_clean', converts the time filed to New Zealand Standard Time (NZST) and removes any duplicates present within the data. There is also code included to check that any new data added is being processed correctly. Next, '02_preprocess' is ran, creating data tables needed for computation and joining together data required for the dataset exportation. This section also contains graphs in order for data to be visually examined. The next script, '03_compute', performs the computations that provide the population estimate. The final script, '04_visualization', produces the final visualizations and exports the dataset that we were commissioned to produce.

Methods

In this section, we explain the methods used in the pipeline to produce our outputs.

Population Estimate

We have used a basic method for estimating the population that involves just two calculations. The first of these estimates the population of people with devices at all times by flattening the total number of devices to equal the maximum number of devices at all times. This estimate is necessary due to large fluctuations in the number of devices throughout the day, as shown in Figure 1.

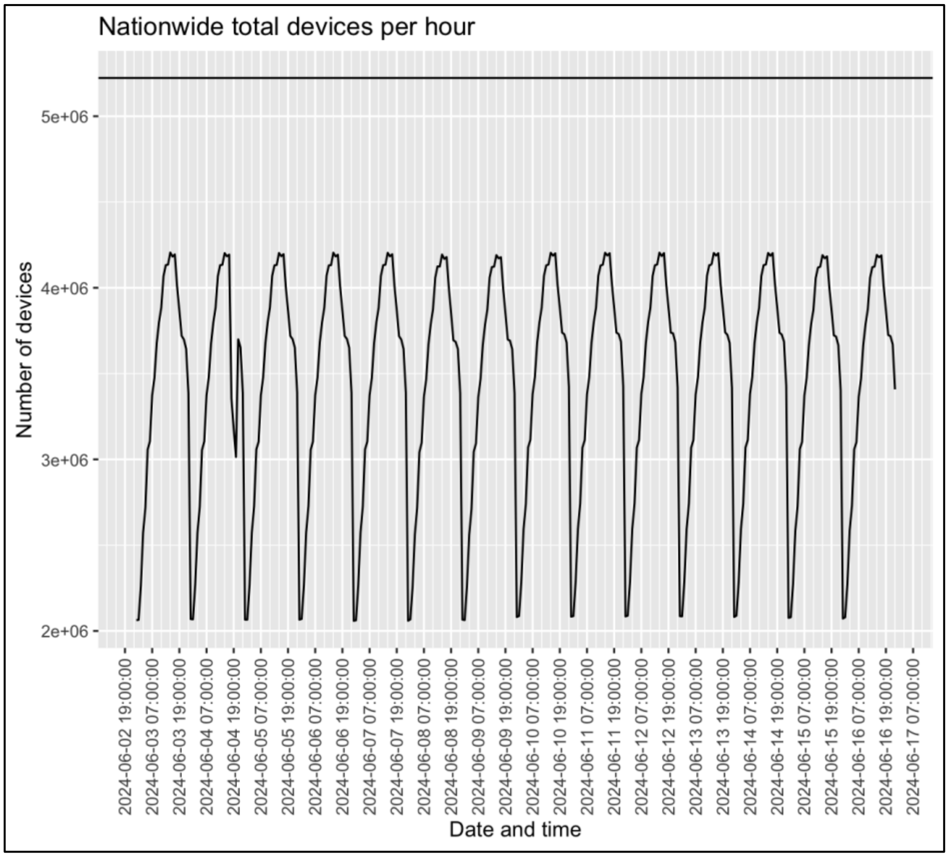

Figure 1: Line graph that shows the total number of devices in the whole country for every hour over the two-week window. There is also a horizontal line that shows the total population of the country.

The second calculation estimates the number of people that don’t have devices. In the creation of its formula, we have made a few assumptions. The first of which being that every person only has one phone. The second being that the maximum hour in the dataset has every device connected, thus, all people in the population that have devices are counted at that time. The third assumption is that the total number of devices in the country stays constant during this time, and that any decrease in device count is due to the device not being included in the counts.

Our method performs the calculation in two steps. The first of which takes the hour with the most devices counted in the entire country and divides it by each hour's device count (for the whole country), creating a multiplier called ‘formula’. The purpose of this multiplier is to increase the number of devices in each region to match the count of the maximum hour which we assumed is the total number of devices. To do this, we take the multiplier 'formula', which changes every hour, and multiply it against the number of devices in each region. This provides us with a final estimate of the number of people with devices.

We then need to estimate the number of people without devices. To do this we make another multiplier, 'offset_multiplier', created by taking the total population and dividing it by the hour with the most devices. To obtain the population without devices estimate, the estimated population from the previous step is multiplied by 'offset_multiplier'. The resulting estimate, seen in Figure 2, is now a constant for all hours of the day relative to the total population estimate for the country.

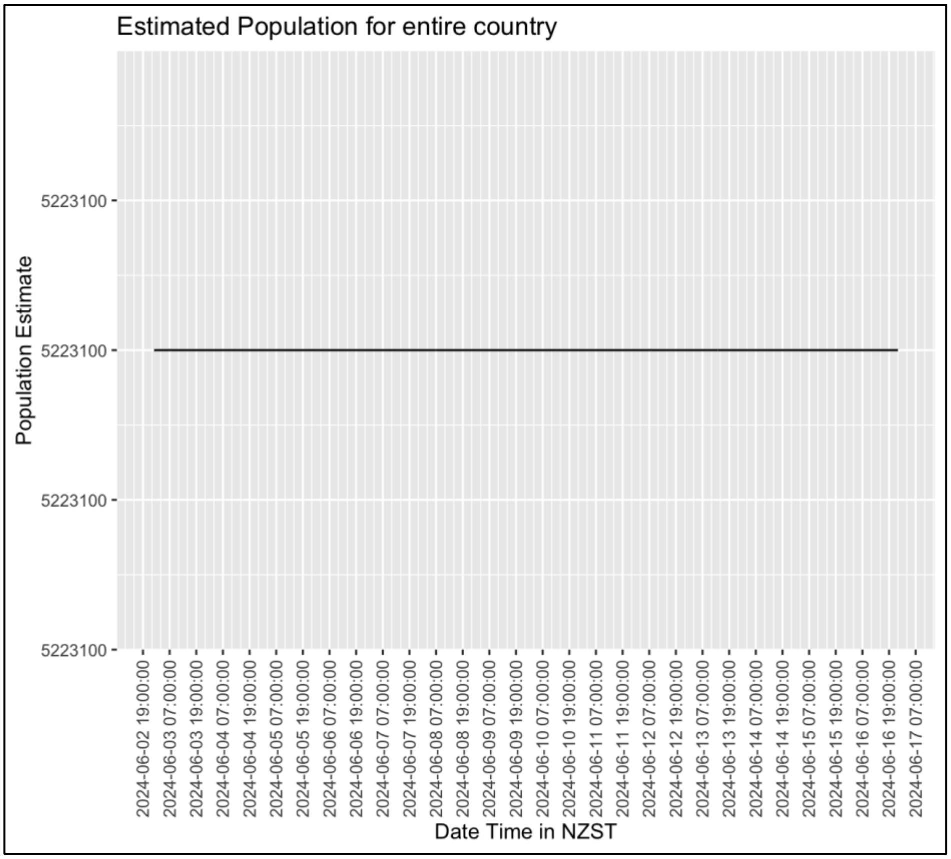

Figure 2: Population Estimate graph shows how after the conversion the population per hour is flat.

Pseudo Code

```
ts = timestamp
Formula[ts] = max_hour_devices/current_hour_devices
Estimated_population_with_devices[ts, sa2] = formula * sa2_device_count_current_hour
Offset_multiplier = country_populaton / max_hour_devices
Estimated_population[ts, sa2]  = Estimated_population_with_devices * offset_multiplier
```

Conversions

In the process of making the dataset, multiple conversions had to be made. The first is the conversion of the spark telecommunications dataset from UTC to NZST to properly fit our requirements. The ‘vf’ (Vodafone) dataset also had the date/time formatted to NZST but looked to already be in the correct format, just not labelled as such.

Corrections

Removed duplicate entries that were fully identical.

Removed entries with the same region and timestamp, taking the one with the highest value.

Limitations

- Not everyone has a phone, and some people can have more than one. As we are tracking populations from data collected by telecommunication companies, this could cause an imbalance in our estimations.
- We would expect that the number of individuals with more than one phone will be leaving at the end of the workday, as we are presuming that they have both a personal phone and a work phone.
- Our dataset only covers a 2-week period (one week school holidays; one not). This could potentially mean that our results might not be an accurate reflection of the population.
- Not everyone using a phone in each region may be a road user. For example, they might be someone living there and is staying home all day or might have other options/preferred travel methods that wouldn’t be greatly affected by roading infrastructure upgrades. The data also doesn't indicate what is necessarily traffic data as the data provided does not indicate if anyone in these regions is moving or potentially just being disconnected from the network.

- There is no differentiation in these datasets between mobile phones and any fixed-point devices that also connect to the mobile networks like many IOT devices available. Therefore, we have had to assume that all the devices in our counts were mobile phones, but this could be incorrect.

Future Improvements

- A larger dataset (longer than 2 weeks) could greatly improve the accuracy of our claims.
- It would be useful to utilise an additional dataset to help correlate the telecommunication data to actual road users, such as road-based sensors. This could be used to correlate actual traffic to the number of devices in the area.
- Could do more research around the average number of devices per person and whether that changed depending on the area. For example, how many individuals have devices within urban and rural areas, or if the number of devices connected per hour was influenced using work phones being switched off at certain times.
- Potentially investigate a method that uses territorial population estimates instead of the whole country as this might generalise better for smaller regions.

CBD Definition

- The Auckland CBD SA2s were defined using Auckland Council's 2024 City Centre Local Alcohol Policy map (https://www.aucklandcouncil.govt.nz/plans-projects-policies-reports-bylaws/our-policies/Documents/city-centre-lap-map.pdf) and matched with their respective SA2s using Statistics NZ's generalised Statistical Area 2 2023 map (https://datafinder.stats.govt.nz/layer/111227-statistical-area-2-2023-generalised/). The defined SA2s were then cross checked with Auckland Council's 2018 Census results report (https://www.censusauckland.co.nz/files/Auckland%20City%20Centre%202018%20Census%20info%20sheet.pdf).
- The Wellington CBD SA2s were defined by using the 2023 Wellington City Council Housing and Business Needs Assessment (HBA) report map (https://wrlc.org.nz/wp-content/uploads/2024/04/HBA3-CHAPTER-2-Wellington_16.02.24.pdf) showing housing catchment areas in Wellington city (page 95 of the report) and mapping the Wellington Central/CBD area into SA2s using Statistics NZ's generalised Statistical Area 2 2023 map. The SA2s were also cross checked with the Wellington CBD SA2s defined by Stats NZ's in the 2018 census (https://www.stats.govt.nz/news/newly-released-census-data-shows-christchurch-cbd-bouncing-back).

- Christchurch City Council's Central City neighbourhood map (https://ccc.govt.nz/culture-and-community/central-city-christchurch/live-here/our-central-neighbourhoods) and Statistics NZ's generalised Statistical Area 2 2023 map were collectively utilised to define Christchurch CBD's SA2s. The defined SA2s for Christchurch CBD was also checked with Stats NZ's defined Christchurch CBD SA2s in the 2018 census (https://www.stats.govt.nz/news/newly-released-census-data-shows-christchurch-cbd-bouncing-back).

Limitations to defining CBDs

- Factors such as urban development, economic shifts and demographic changes can cause CBDs to change over time, therefore, fixed definitions of CBD SA2s may not thoroughly capture these changes.

Future Refinements to defining CBDs

- Models that can incorporate real-time data on aspects that define CBDs could be developed to ensure that CBDs are defined with data that accurately reflects the areas.
- Include community and stakeholder's input to refine definitions e.g. feedback from local businesses/residents.
- Regular monitoring and evaluation of CBDs, to ensure they accurately reflect the current conditions of the cities.